Development of an Italian RM Y-STR haplotype database: Results of the 2013 GEFI collaborative exercise

Development of an Italian RM Y-STR haplotype database: results of the 2013 GEFI collaborative exercise

Robino C[a*], Ralf A[b], Pasino S[a], De Marchi MR[a], Ballantyne KN[c], Barbaro A[d], Bini C[e], Carnevali E[f], Casarino L[g], Di Gaetano C[h,i], Fabbri M[j], Ferri G[k], Giardina E[l], Gonzalez A[m], Matullo G[h,i], Nutini AL[n], Onofri V[o], Piccinini A[p], Piglionica M[q], Ponzano E[r], Previderè C[s], Resta N[t], Scarnicci F[u], Seidita G[v], Sorçaburu-Cigliero S[w], Turrina S[x], Verzeletti A[y], and Kayser M [b].

[a] Department of Public Health Sciences and Pediatrics, University of Turin, Italy

[b] Department of Forensic Molecular Biology, Erasmus MC University Medical Center Rotterdam, The Netherlands

[c] Office of the Chief Forensic Scientist, Victoria Police Forensic Services Department, Macleod, Australia

[d] Department of Forensic Genetics, Studio Indagini Mediche e Forensi (SIMEF), Reggio Calabria, Italy

[e] Department of Medical and Surgical Sciences, Institute of Legal Medicine, University of Bologna, Italy

[f] Department of Biomedical and Surgical Sciences, Section of Legal Medicine and Forensic Science, University of Perugia, Italy

[g] Dipartimento di Medicina Legale, del Lavoro, Psicologia Medica e Criminologia, Università di Genova, Italy

[h] Department of Medical Sciences, University of Turin, Italy

[i] HuGeF, Human Genetics Foundation, Turin, Italy

[j] Department of Public Health, UOL of Legal Medicine, University of Ferrara, Italy

[k] SC Medicina Legale, Università di Modena, Italy

[l] Department of Biomedicine and Prevention, University of Rome "Tor Vergata", Italy

[m] ANDROS Day Surgery Clinic, Forensic Genetics Unit, Palermo, Italy

[n] SOD Genetics Diagnostics, Forensic Genetics, Azienda Ospedaliera Universitaria Careggi, Florence, Italy

[o] Section of Legal Medicine, Università Politecnica Delle Marche, Ancona, Italy

[p] Dipartimento di Scienze Biomediche per la Salute, Università degli Studi di Milano, Italy

[q] Interdisciplinary Department of Medicine, Section of Legal Medicine, University of Bari, Italy

[r] Department of Molecular Medicine, University of Padova, Italy

[s] Department of Public Health, Experimental and Forensic Medicine, University of Pavia, Italy

[t] Department of Biomedical Sciences and Human Oncology, Medical Genetics Unit, "Aldo Moro" University of Bari, Italy

[u] Istituto di Medicina Legale, Università Cattolica del Sacro Cuore, Roma, Italy

[v] Department of Biopathology, Medical and Forensic Biotechnologies, University of Palermo, Italy

[w] Department of Medicine, Surgery and Health, University of Trieste, Italy

[x] Dipartimento di Sanità Pubblica e Medicina di Comunità, Università degli Studi di Verona, Italy

[y] Department of Medical and Surgical Specialties, Radiological Sciences and Public Health, University of Brescia, Italy

*** Corresponding author:** Department of Public Health Sciences and Pediatrics, University of Turin, Corso Galileo Galilei 22, 10126 Turin, Italy

e-mail: carlo.robino@unito.it (C. Robino)

Abstract

Recently introduced rapidly mutating Y-chromosomal short tandem repeat (RM Y-STR) loci, displaying a multiple-fold higher mutation rate relative to any other Y-STRs, including those conventionally used in forensic casework, have been demonstrated to improve the resolution of male lineage differentiation and to allow male relative separation usually impossible with standard Y-STRs.

However, large and geographically-detailed frequency haplotype databases are required to estimate the statistical weight of RM Y-STR haplotype matches if observed in forensic casework. With this in mind, the Italian Working Group (GEFI) of the International Society for Forensic Genetics launched a collaborative exercise aimed at generating an Italian quality controlled forensic RM Y-STR haplotype database. Overall 1509 male individuals from 13 regional populations covering northern, central and southern areas of the Italian peninsula plus Sicily were collected, including both "rural" and "urban" samples classified according to population density in the sampling area. A subset of individuals was additionally genotyped for Y-STR loci included in the Yfiler and PowerPlex Y23 (PPY23) systems (75% and 62%, respectively), allowing the comparison of RM and conventional Y-STRs. Considering the whole set of 13 RM Y-STRs, 1501 unique haplotypes were observed among the 1509 sampled Italian men with a haplotype diversity of 0.999996, largely superior to Yfiler and PPY23 with 0.999914 and 0.999950, respectively. AMOVA indicated that 99.996% of the haplotype variation was within populations, confirming that genetic-geographic structure is almost undetected by RM Y-STRs. Haplotype sharing among regional Italian populations was not observed at all with the complete set of 13 RM Y-STRs. Haplotype sharing within Italian populations was very rare (0.27% non-unique haplotypes), and lower in urban (0.22%) than rural (0.29%) areas. Additionally, 422 father-son pairs were investigated, and 20.1% of them could be discriminated by the whole set of 13 RM Y-STRs, which was very close to the theoretically expected estimate of 19.5% given the mutation rates of the markers used. Results obtained from a high-coverage Italian haplotype dataset confirm on the regional scale the exceptional ability of RM Y-STRs to resolve male lineages previously observed globally, and attest the unsurpassed value of RM Y-STRs for male-relative differentiation purposes.

Keywords: Y-chromosome; rapidly mutating Y-STRs (RM Y-STRs); haplotype; lineage differentiation; relative differentiation; Italy

1. Introduction

Analysis of Y-chromosomal short tandem repeat (Y-STR) loci provides an extremely useful tool in forensic DNA testing. In particular, it allows the unambiguous detection of the male DNA component in mixtures with a high female background, as often found in sexual assault cases. Moreover, due to their haploid nature and uniparental transmission favouring the geographical clustering of haplotypes, Y-STRs can provide intelligence information on the ethnic origin of a stain donor in non-suspect cases. The principal weakness of Y-STR analysis is that, even when a crime sample matches the Y-STR haplotype of a suspect, his patrilineal relatives cannot be excluded as being the donor of the stain [1]. Adding additional markers to the current sets of Y-STRs used in forensic casework can improve the level of paternal lineage differentiation [2]. Since mutation is the only genetic force behind Y-haplotype variation, Y-STRs displaying high mutation rates are best fitted for this purpose. A systematic Y-STR mutation study by Ballantyne et al. [3] identified a set of 13 novel "rapidly mutating" (RM) Y-STR markers with exceptionally high mutation rates ($>10^{-2}$ per locus per generation) if compared to >170 other Y-STRs including all conventionally used markers, the latter in the order of a few mutations per marker every 1,000 generations [4]. RM Y-STRs have been demonstrated not only to improve the resolution of male lineage differentiation, but also to allow close male relative separation with a power unsurpassed by standard Y-STRs [5].

Estimation of the statistical weight of Y-STR haplotype matches is further complicated by the fact that Y chromosomes are highly geographically structured, requiring adequate reference databases -in terms of size and regional coverage- to reflect the population-wide spectrum of Y-STR haplotypes in sufficient detail. A recent survey of the variability of the 13 RM Y-STRs across 111 worldwide populations [6], while confirming their unequalled value for male lineage differentiation with <99% of the >12,200 unrelated men being completely individualized, also indicates that the extremely high mutation rates of RM Y-STRs almost erase any signal of population substructure, at least at a global level. As a consequence it was suggested that the need for regional (metapopulation) reference databases for haplotype frequency estimation in forensic applications is strongly reduced for this RM

Y-STR set, compared to standard Y-STR sets [6]. Given the global nature of this recent study [6], it would be interesting to investigate RM Y-STR haplotype diversity and distribution within a geographic region. In this respect, an extensive analysis of RM Y-STR diversity in Italy may prove itself especially revealing. Due to its central position in the Mediterranean sea, Italy has historically been a convenient destination for human populations migrating from Africa, the Middle East and European locations; this, along with its varied and rugged geomorphological characteristics, have contributed to shape a complex mosaic of genetic variation. On the Y-chromosomal perspective, though a North-South major cline across the Italian Peninsula was described, local drift and founder effect had been signalled as the main explanation for the observed distribution of genetic diversity [7].

With all this in mind, the Italian Working Group (GEFI) of the International Society for Forensic Genetics (ISFG) organized a collaborative exercise focused on RM Y-STRs. The study was aimed at the implementation of these new markers in member laboratories, based on internal validation of the proposed typing protocol through quality control procedures. In order to start establishing a reference database for haplotype match calculations, RM Y-STR variation in Italy was then investigated by adopting an effective sampling strategy, that combined wide national coverage and high resolution, including individuals from both rural and urban areas of the peninsula.

2. Materials and methods

2.1. DNA sample collection

Detailed geographic localization and composition of the tested populations is displayed in Figure 1. A total of 1509 DNA samples, obtained from consenting adult males originating from thirteen Italian regional populations defined by political boundaries, were analyzed. Regional population sample sets ranged in size between 24 and 304 individuals (median 83 individuals per population sample). Collected samples were categorized as “rural” and “urban”, according to population density in the sampled area [8]. Sampling strategy implied the exclusion of known close (i.e. first- and second-

degree) relatives. Individuals from rural areas were carefully selected based on genealogical data, in order to include only subjects with at least three generations of residence in the sampling area. In order to test the ability of RM Y-STRs to differentiate between close male relatives, 422 father-son pairs previously confirmed by autosomal DNA analysis were also investigated.

2.2. Y-STR genotyping

The 13 RM Y-STR markers were amplified in three multiplex PCR assays: RM1 (DYS576, DYS570, DYF387S1 and DYF399S1); RM2 (DYS626, DYS627, DYS526, and DYS518); RM3 (DYS612, DYF404S1, DYS449, DYS547 and DYF403S1). Primer sequences were according to [3], with the only exception of locus DYS612 [6]. Novel forward primers for locus DYS576 (5'-GTTGGGCTGAGGAGTTCAATC-3') and DYS404S1 (5'- TGGCAGGACACATTTAAACA -3') were used. Dye-labeling and PCR primers concentration were as described in [6]. Template DNA (0.5-2 ng) was amplified in 10 µl PCR reactions, containing 1 X Qiagen PCR Master Mix (Qiagen, Hilden, Germany). All the three multiplex PCR reactions were performed using a single touchdown protocol (94 °C for 10 min, 10 cycles of 94 °C for 30 s, 65–1 °C every cycle for 30 s and 72 °C for 1 min, followed by 25 cycles of 94 °C for 30 s, 50 °C for 30 s and 72 °C for 1 min, with 45 min at 60 °C). PCR amplified products were then separated and detected using participating laboratories' standard capillary electrophoresis (CE) protocols for analyzing STRs, in either ABI310, ABI3130, or ABI3500 Genetic Analyzers with POP-4, POP-6 or POP-7 (Life Technologies, Foster City, CA), and size standard ILS-600 (Promega Corporation, Madison, WI). Allele calling was performed with GeneMapper software (Life Technologies, Foster City, CA) utilizing custom panel and bin sets, as well as allelic ladders as described elsewhere [6].

A subset of the population sample was also typed with the AmpFlSTR Yfiler PCR Amplification Kit (Yfiler, Life Technologies, Foster City, CA) (n=1133), and the PowerPlexY23 System (PPY23, Promega Corporation, Madison, WI) (n=938), according to the manufacturer's instructions. All Yfiler/PPY23 haplotype data used in the study are publicly available at the Y Chromosome Haplotype

Reference Database (YHRD) website (www.yhrd.org), in compliance with the guidelines for publication of population data requested by the journal [9]. YHRD accession numbers and size of each population sample typed with Yfiler/PPY23 are given in Supplementary material, Table S1.

2.3. Quality control

To ensure genotyping consistency between the laboratories, all participants received two sets of four blind control DNA samples as used in [6]. Haplotypes of control DNAs 007 [6] and 2800M (Supplementary material, Table S2) included in Yfiler and PPY23 kits, respectively, were also provided for ladder calibration. Nomenclature of RM Y-STR alleles was according to [6], in compliance with ISFG guidelines [10].

Genotyping of population samples was only allowed after a participating laboratory demonstrated the correct genotyping of the first set of blind control DNA samples at all 13 RM Y-STRs. In case participants reported erroneous genotypes, screenshots were requested and submitted to the organizing laboratory (Department of Public Health Sciences and Pediatrics, University of Turin) for evaluation. After the identification of the possible cause of error, laboratories were requested to type the second set of blind control DNAs. In case this was done correctly, they were allowed to type their population samples. Submitted population samples with missing data from more than one marker were excluded from data analysis, to prevent low quality samples affecting genotype and haplotype distributions. Any differences observed between father-son pairs were confirmed through duplicate, independent PCR amplifications and genotyping.

2.4. Statistical analysis

Statistical calculations of standard diversity indexes (haplotype diversity, *h*; mean number of differing loci), pairwise genetic distances (F_{ST}) and analysis of molecular variance (AMOVA) were performed with the software Arlequin version 3.5.1.2 [11]. Estimation of the coancestry coefficient θ in the tested populations was done according to [12].

Multidimensional scaling (MDS) analysis [13] was performed using the software XLSTAT (Addinsoft, Paris, France). MDS analyses of population matrixes of Slatkin's linearized F_{ST} values were performed for one to 10 dimensions. Optimal dimensionality was obtained iteratively reducing Kruskal's stress value until it remained nearly unchanged.

Testing of statistical significance (z-test, Fisher exact test, t-test) was performed using the software SigmaStat v3.2 (Systat software, San Jose, CA).

3. Results

3.1. Quality control

Twenty-one GEFI laboratories returned results for the first set of four blind control DNA samples. Review of the data combined with inspection of electropherograms, while revealing sporadic clerical errors and a few cases of bin misalignment, which were quickly identified and corrected, showed that all remaining errors involved the multi-copy markers DYF399S1 and DYF403S1. In particular 11 laboratories reported incorrect genotypes for one (91%) or two (9%) blind control DNA samples at locus DYF399S1. Locus DYF399S1 includes microvariant alleles separated by just one base pair, which can be difficult to differentiate under poor CE resolution conditions. Moreover ambiguous allele calls may depend on PCR artifacts, like non template addition in the presence of excess template DNA [14]. Participants were therefore advised to strictly control the amount of input DNA in PCR reactions for multiplex RM1, and to switch to higher concentrated sieving polymers (as POP-6 and POP-7) for CE. Among the 11 aforementioned laboratories, 6 laboratories also incurred in one (33%) or two (66%) errors when typing DYF403S1, principally caused by the occurrence of extra peak artifacts (> 340 bp) at locus DYF403S1a. However, such artifacts could be effectively eliminated after the adoption for multiplex RM3 of conventional PCR conditions (35 amplification cycles with annealing at 60 °C 30 s), instead of the originally suggested touchdown PCR protocol. Given these indications, all laboratories were able to provide correct genotypes for the second set of four blind control DNA samples, and consequently allowed to participate in the population study. However, due

to concerns raised during the quality control process regarding the genotyping accuracy of loci DYF399S1 and DYF403S1 [6], all following calculations were performed both considering the whole set of markers (13 RM Y-STRs), and excluding DYF399S1 and DYF403S1 (11 RM Y-STRs).

3.2. Diversity of RM Y-STR haplotypes

Complete RM and conventional Y-STR haplotypes observed in the Italian population sample are listed in Supplementary material, Table S3. Standard diversity indexes for 13 and 11 RM Y-STRs in the total set of 1509 individuals, summarized according to regional populations and rural/urban origin, are shown in Table 1. Considering the whole set of 13 RM Y-STRs, the haplotype diversity estimate was h=0.999996 with 1505 different haplotypes among 1509 individuals tested. At the regional population level, similarly high levels of haplotype diversity were observed across all populations, ranging from h=1, i.e. complete individualization with only unique haplotypes observed in 85% of the populations tested, to h=0.999935 in Sicily (301 different haplotypes among 304 individuals tested). In general haplotype sharing was very rare (0.27% non-unique haplotypes) and lower in urban (0.22%) than rural (0.29%) areas; this difference however was not statistically significant (z=-0.292, p=0.770). Notably, no haplotype sharing between regional populations was observed. The four non-unique haplotypes were shared between individuals from the same regional population i.e., one pair in a single rural village of Latium (Collevecchio), two pairs in the rural Sicilian village of Santa Ninfa, and one pair in the Sicilian urban area of Trapani. Detailed geographic localization of sampling sites is displayed in Supplementary material, Figure S1.

The resolution of male lineages in Italy was only slightly reduced by the exclusion of DYF399S1 and DYF403S1 (11 RM Y-STRs), with h=0.999983 in the total dataset (1490 haplotypes among 1509 males). In this case 19 haplotypes were observed twice, 14 pairs in rural areas, 4 pairs in urban areas, and one pair between rural and urban areas. Frequency of shared haplotypes was lower in urban (1.09%) than rural (1.45%) areas, though without reaching statistical significance (z=0.315, p=0.753). All haplotype matches were within regional populations, with the only exception of one haplotype

pair which was shared between two urban individuals from neighbouring Piedmont (Cuneo) and Lombardy (Pavia) (Figure S1). However, differential distribution of interregional matches between rural and urban populations did not reach statistical significance (Fisher exact test, p=0.078).

Comparison with 13 RM Y-STR haplotypes observed in 12,072 samples from 111 worldwide populations [6], after excluding 200 Italian samples reported both in [6] and in the present dataset, did not detect any match in the Italian population studied here. For 11 RM Y-STR haplotypes, one single match was observed between a rural sample from Abruzzo (Montereale) in the present study and Austria (Salzburg) in the worldwide study [6] (Figure S1). In Table 2 standard diversity indexes are compared in a subset of Italian individuals who were genotyped for both RM and conventional Y-STRs included in the Yfiler (n=1133) and PPY23 (n=938) systems. It can be seen that both the number of haplotypes and haplotype diversity values were constantly higher (and θ values lower) with the RM Y-STRs, even when disregarding the two multi-copy loci DYF399S1 and DYF403S1. Only in the urban subsample, the 11 RM Y-STR subset and PPY23 showed equal power of resolution (*h*=0.999906) of male lineages. One single non-unique 13 RM Y-STR haplotype (Santa Ninfa) could be further discriminated by conventional Y-STR kits, through a single mismatch at locus DYS635.

The mean number of differing RM loci in pairs of individuals sharing non-unique Yfiler and PPY23 haplotypes was 5.5 ± 4.5 SD and 3.5 ± 3.5 SD, respectively. In pairs sharing the same Yfiler haplotype, the mean number of mismatches was significantly lower (t_{53}=6.794, p>0.001) in subjects having the same regional background (4.4 ± 3.7 SD) compared to those originating from different regions (12.4 ± 3.3 SD). Though only one PPY23 haplotype pair was shared between two subjects from different regional populations, thus preventing meaningful statistical comparisons, also in this case the number of mismatches found at RM Y-STR loci (15) was strikingly higher than the average number observed in pairs of individuals with identical PPY23 haplotype from the same region (3.0 ± 2.6 SD). The proportion of individuals carrying non-unique haplotypes which were shared among regions was significantly larger in urban rather than rural populations considering PPY23 (Fisher exact test, p=0.032), though not for Yfiler (Fisher exact test, p=0.322).

Since forensic DNA testing often requires the analysis of degraded DNA, with a pronounced drop-out of longer STR amplicons, haplotype diversity calculations were also performed considering -for each Y-STR panel- only the loci with amplicon lenght approximately lower than 250 bp, that is: DYS576, DYS526a, DYS626, DYS612 and DYF404S1 (RM Y-STRs); DYS456, DYS389I, DYS390, DYS458, DYS19, DYS393, DYS391, DYS439, YGATAH4, DYS437 and DYS438 (Yfiler); DYS576, DYS389I, DYS448, DYS391, DYS481, DYS549, DYS570, DYS635, DYS390, DYS393 and DYS458 (PPY23). In this analysis, RM Y-STRs, in spite of the limited number of eligible loci (5), resulted superior to Yfiler (11 loci used) with haplotype diversity values of h=0.9997 and h=0.9996, respectively. PPY23 (11 loci used) showed a higher power of resolution (h=0.9998). However it must be noted that in this case PPY23 took advantage from the inclusion of the RM locus DYS570, which had to be dropped from calculations in the RM Y-STR panel, due to the different primer design used in the present study that generated >250 bp amplicons. If also DYS570 were considered in the eligible subset of short amplicon RM Y-STR markers, the value of haplotype diversity that could be reached (h=0.9998) would then equal that obtained with PPY23.

3.3. Genetic structure and population comparisons

Considering both the whole set of 13 RM Y-STRs and the 11 RM Y-STR subset, overall and regional θ values were extremely low (Table 1), confirming that relatively little population substructure is detected with RM Y-STRs. Therefore, no specific correction seems necessary when using these markers in forensic applications, as is usually needed for other DNA polymorphisms [15].

For AMOVA analysis, Italian regional population samples were grouped according to their geographic position (North: Piedmont, Lombardy, Veneto, Friuli Venezia Giulia, Emilia Romagna; Central: Tuscany, Umbria, Marche, Latium, Abruzzo; South: Apulia, Calabria, Sicily). Calculations were based on F_{ST} values, since the inability to accurately assign alleles of multi-copy markers to specific loci and the presence of frequent microvariant alleles in several RM markers prevented R_{ST} values being used. AMOVA results (Table 3) demonstrated that for 13 RM Y-STR haplotypes

99.996% of the observed variation was within populations, 0.003% among populations within groups, and 0.001% among groups, in line with what previously found at a global level [6]. The proportion of variance ascribable to haplotype differences among groups remained negligible in the 11 RM Y-STR subset (0.000%), raising to 0.005% for Yfiler and PPY23. Both in RM and conventional Y-STRs variation among populations within groups was constantly higher in the rural rather than in the urban subsample. Population pairwise F_{ST} distances were extremely low for RM Y-STRs, with average values of 0.00003 (13 RM loci) and 0.00013 (11 RM loci). Again, average pairwise F_{ST} values were larger between rural (0.00013, 13 RM Y-STR; 0.00022, 11 RM Y-STR) rather than urban populations (0.00002, 13 RM Y-STR; 0.00015, 11 RM Y-STR). The difference in average F_{ST} values between rural and urban populations was significant for the complete 13 RM loci set (t_{121}=2.467, p=0.015), though not for the 11 RM loci set (t_{121}=1.172, p=0.243). These results further highlight the role of local drift and founder effect previously invoked in shaping the Y-chromosomal landscape of Italy [7,16].

For all Y-STR panels (RM Y-STRs, Yfiler, PPY23) two MDS components proved optimal to summarize Y-chromosomal diversity in Italian populations, based on linearized F_{ST} distances (Figure 2). No defined geographic pattern emerged when regional populations were analyzed by means of RM Y-STRs. On the contrary, some North-South discontinuity appeared for PPY23 loci, and became clearly evident in Yfiler loci. In this last case it is particularly interesting to observe how the regional sample from Veneto is positioned closer to Central Italian populations, in strict concordance with a recent extensive study of uniparental markers in Italy, which showed that Y-chromosomal variation is arranged along a North West – South East, rather than North-South, axis along the peninsula [17].

3.4. Male relative differentiation with RM Y-STRs

Eighty-five father-son pairs out of 422 (20.1%) could be discriminated by at least one RM Y-STR allele mismatch. This empirically obtained percentage, though slightly lower than that observed in a previous larger study of 2378 father-son pairs (26.9%) [6], is very close to the theoretical estimate

(19.5%) based on the average mutation rate of the set of RM Y-STR markers obtained from the sampling from the posterior distribution [3]. Of the 85 resolved father-son pairs, 89.4% showed allele discrepancies at one locus, and 10.6% at two loci. The large majority of the observed mutations (90.4%) consisted of the loss/gain of one repetitive unit, in general concordance with the classical step-wise mutation model of STR evolution [18]. The overall average mutation rate was estimated to be 1.75×10^{-2} (95% CI $1.4 \times 10^{-2} – 2.1 \times 10^{-2}$), very close to the estimate by Ballantyne et al. [3] of 1.97×10^{-2} (95% CI $1.8 \times 10^{-2} – 2.2 \times 10^{-2}$).

It must be stressed that after excluding from calculations the multi-copy markers DYF399S1 and DYF403S1, the power of resolution of father-son pairs dramatically dropped to 9.7%. This clearly indicates how, in spite of possible concerns regarding the robustness of genotyping, inclusion of these loci in the RM Y-STR panel is crucial for this specific application.

4. Discussion

The GEFI collaborative study on RM Y-STRs allowed member laboratories to implement their current panel of markers used in forensic testing and genealogical studies with a new set of polymorphic loci with unsurpassed power of male lineage separation and male relative differentiation. Generally speaking, the future application of these markers in forensic casework poses no different problems than those raised by the analysis of autosomal STR profiles in stains, namely the possibility of allele drop-out and/or drop-in, and the occurrence of mixtures [19,20]. However, such issues have not been fully explored yet in haploid markers [21]. In particular the deconvolution of mixed genotypes of multi-copy RM Y-STR markers with non fixed number of alleles, whose adoption appears all-important for male relative discrimination purposes, may prove particularly challenging based on available interpretation models [22,23], and will require the development of new ad hoc statistical solutions.

The present study also highlighted how future optimization of current RM Y-STR PCR and CE protocols is feasible, possibly through the adoption of either 5-dye or newly available 6-dye

chemistries [24]. Such is the case of locus DYS570, which in PPY23 is amplified by means of PCR primers giving rise to amplicons which are decidedly shorter (90-146 bp) [25] than those obtained with the present protocol (246-286 bp) [5], and therefore less prone to the potential effects of DNA degradation.

As previously stated, forensic interpretation of Y haplotype matches requires large and detailed population reference databases. In this respect, the GEFI study allowed to collect extensive information on RM Y-STR variation in the Italian peninsula, with an over 7.5-fold increase of the Italian RM Y-STR haplotype database firstly made available through the worldwide survey carried out by Ballantyne et al. [6]. It was shown that, even when restricting the analysis to a single European country, the vast majority of the observed haplotypes were singletons (99.5% and 97.5% for the 13 and the 11RM Y-STR panels, respectively). This means that in forensic casework most suspect-stain RM Y-STR matches will likely involve haplotypes previously undetected in the reference database, a circumstance under which traditional count estimates of the corresponding match probabilities become unsatisfactory. Several contrasting approaches have been proposed to empirically derive the frequency estimates of such rare Y-STR haplotypes [26-29]. For estimators like "frequency surveying" [28], that try to draw further information from the evolutionary relatedness of constituent haplotypes in the reference database, the obtained results can provide an useful framework for the interpretation of RM Y-STR matches.

The almost complete absence of population substructure detected for RM Y-STR loci, previously observed at a global level [6], was confirmed within a single geographic region, like Italy, for which clinal patterns of Y chromosomal variation had been extensively documented [7,16,17]. This property of RM Y-STRs reduces the need for regional (metapopulation) reference databases for haplotype frequency estimation compared to standard Y-STR sets. On the other hand, our data also show that the uncertainty about the true haplotype frequency is larger in rural areas where more related men are expected to be found than in urban areas. As a consequence, in order to reflect the amount of male population substructure in a region, it is advisable that Y-STR forensic reference databases -unlike

those developed for anthropological purposes- should be derived from a random set of subjects, including related and unrelated individuals [30].

Finally, our study demonstrated that RM Y-STRs not only can effectively dissect male lineages unresolved by conventional Yfiler/PPY23 loci, but also may be able to provide investigators with important additional leads. Subjects sharing the same Yfiler/PPY23 haplotype showed a significantly higher number of allelic mismatches at RM Y-STR loci when from different Italian regional populations, compared to individuals from the same region. It is reasonable to explain the occurrence of interregional Yfiler/PPY23 haplotype sharing observed in our dataset as the consequence of recurrent mutation, whereas relatedness is most likely for matches within regional populations, especially in rural sampling sites. Presence/absence of RM Y-STR near matches in samples displaying the same conventional Y-STR haplotype may therefore be helpful to discriminate between adventitious Yfiler/PPY23 matches, arising in individuals that are not of common descent, and truly related subjects. Though already observed for PPY23 loci compared to Yfiler [31], this effect reaches its maximum when RM Y-STRs are considered, so that observation of a RM Y-STR near match between the stain donor and a suspect would definitely support the need for familial searching among the male relatives of the alleged offender, or for a DNA dragnet in the specific geographical area of origin of his male ancestors.

References

1. L. Roewer L, Y chromosome STR typing in crime casework, Forensic Sci. Med. Pathol. 5 (2009) 77-84.
2. J. Purps, S. Siegert, S. Willuweit, M. Nagy, C. Alves, et al., A global analysis of Y-chromosomal haplotype diversity for 23 STR loci, Forensic Sci. Int. Genet. 12 (2014) 12-23.

3. K.N. Ballantyne, M. Goedbloed, R. Fang, O. Schaap, O. Lao, et al., Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications, Am. J. Hum. Genet. 87 (2010) 341-353.
4. M. Goedbloed, M. Vermeulen, R.N. Fang, M. Lembring, A. Wollstein, et al., Comprehensive mutation analysis of 17 Y-chromosomal short tandem repeat polymorphisms included in the AmpFlSTR Yfiler PCR amplification kit, Int. J. Legal Med. 123(2009) 471-482.
5. K.N. Ballantyne, V. Keerl, A. Wollstein, Y. Choi, S.B. Zuniga, et al., A new future of forensic Y-chromosome analysis: rapidly mutating Y-STRs for differentiating male relatives and paternal lineages, Forensic Sci. Int. Genet. 6 (2012) 208-218.
6. K.N. Ballantyne, A. Ralf, R. Aboukhalid, N.M. Achakzai, M.J. Anjos, et al., Towards male individualization with rapidly mutating Y-chromosomal STRs, Hum. Mutat. 35 (2014) 1021-1032.
7. F. Di Giacomo, F. Luca, N. Anagnou, G. Ciavarella, R.M. Corbo, et al., Clinal patterns of human Y chromosomal diversity in continental Italy and Greece are dominated by drift and founder effects, Mol. Phylogenet. Evol. 28 (2003) 387-395.
8. OECD, OECD Rural Policy Reviews: Italy, OECD Publishing, Paris, 2009.
9. A. Carracedo, J.M. Butler, L. Gusmão, A. Linacre, W. Parson, et al., New guidelines for the publication of genetic population data, Forensic Sci. Int. Genet. 7 (2013) 217-220.
10. L. Gusmão, J.M. Butler, A. Carracedo, P. Gill, M. Kayser, et al., DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis, Int. J. Legal Med. 120 (2006) 191-200.
11. L. Excoffier, G Laval, S. Schneider, Arlequin (version 3.0): an integrated software package for population genetics data analysis, Evol. Bioinform. Online. 1 (2005) 47-50.
12. B.S. Weir, W.G. Hill, Estimating F-statistics Annu. Rev. Genet. 36 (2002) 721-750.
13. J.B. Kruskal, Multidimensional-scaling by optimizing goodness of fit to a nonmetric hypothesis, Psychometrika 29 (1964) 1-27.

14. J.M. Clark JM, Novel non-templated nucleotide addition reactions catalyzed by procaryotic and eucaryotic DNA polymerases, Nucleic Acids Res. 16 (1988) 9677-9686.
15. J.S. Buckleton, M. Krawczak, B.S. Weir BS, The interpretation of lineage markers in forensic DNA testing, Forensic Sci. Int. Genet. 5 (2011) 78-83.
16. C. Capelli, F. Brisighelli, F. Scarnicci, B. Arredi, A. Caglià, et al., Y chromosome genetic variation in the Italian peninsula is clinal and supports an admixture model for the Mesolithic-Neolithic encounter, Mol. Phylogenet. Evol. 44 (2007) 228-239.
17. A. Boattini, B. Martinez-Cruz, S. Sarno, C. Harmant, A. Useli, et al., Uniparental markers in Italy reveal a sex-biased genetic structure and different historical strata, PLoS One. 8 (2013) e65441.
18. A.M. Valdes, M. Slatkin, N.B. Freimer, Allele frequencies at microsatellites loci: the stepwise mutation model revised, Genetics 133 (1993) 737-749.
19. P. Gill, L. Gusmão, H. Haned, W.R. Mayr, N. Morling, et al., DNA commission of the International Society of Forensic Genetics: Recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods, Forensic Sci. Int. Genet. 6 (2012) 679-688.
20. P Gill, C.H. Brenner, J.S. Buckleton, A. Carracedo, M. Krawczak, et al., DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures, Forensic Sci. Int. 160 (2006) 90-101.
21. J. Ge, B. Budowle, R. Chakraborty, Interpreting Y chromosome STR haplotype mixture, Leg. Med. (Tokyo) 12 (2010) 137-143.
22. N. Fukshansky, W Bär, DNA mixtures: biostatistics for mixed stains with haplotypic genetic markers, Int. J. Legal Med. 119 (2005) 285-290.
23. A. Wolf, A. Caliebe, O. Junge, M. Krawczak Forensic interpretation of Y chromosomal DNA mixtures, Forensic Sci. Int. 152 (2005) 209-13.

24. S. Flores, J. Sun, J. King, B. Budowle. Internal validation of the GlobalFiler™ Express PCR Amplification Kit for the direct amplification of reference DNA samples on a high-throughput automated workflow, Forensic Sci. Int. Genet. 10 (2014) 33-39.
25. J.M. Thompson, M.M. Ewing, W.E. Frank, J.J. Pogemiller, C.A. Nolde, et al., Developmental validation of the PowerPlex® Y23 System: a single multiplex Y-STR analysis system for casework and database samples, Forensic Sci. Int. Genet. 7 (2013) 240-250.
26. M.M. Andersen, P.S. Eriksen, N. Morling, The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies, J. Theor. Biol. 329 (2013) 39-51.
27. M.M. Andersen, A. Caliebe, A. Jochens, S. Willuweit, M. Krawczak, Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory, Forensic Sci. Int. Genet. 7 (2013) 264-271.
28. S. Willuweit, A. Caliebe, M.M. Andersen, L. Roewer, Y-STR Frequency Surveying Method: A critical reappraisal, Forensic Sci. Int. Genet. 5 (2011) 84-90.
29. C.H. Brenner CH, Fundamental problem of forensic mathematics--the evidential value of a rare haplotype, Forensic Sci. Int. Genet. 4 (2010) 281-291.
30. M. Kayser, P. de Knijff, Improving human forensics through advances in genetics, genomics and molecular biology, Nat. Rev. Genet. 12 (2011) 179-192.
31. M.H. Larmuseau, N. Vanderheyden, A. Van Geystelen, R. Decorte, A substantially lower frequency of uninformative matches between 23 versus 17 Y-STR haplotypes in north Western Europe, Forensic Sci. Int. Genet. 11 (2014) 214-219

		Haplotypes		*h*		θ		Mean number of differing loci	
	n	13 RM Y-STR	11 RM Y-STR	13 RM Y-STR	11 RM Y-STR	13 RM Y-STR	11 RM Y-STR	13 RM Y-STR	11 RM Y-STR
Piedmont	212	212	210	1	0.999911	0	0.00008612	17.0	11.1
Lombardy	157	157	155	1	0.999837	0	0.00015523	16.9	10.8
Veneto	153	153	153	1	1	0	0	17.5	11.5
Friuli Venezia Giulia	26	26	26	1	1	0	0	17.9	11.8
Emilia Romagna	170	170	167	1	0.999791	0	0.00019464	17.2	11.1
Tuscany	50	50	50	1	1	0	0	17.4	11.3
Umbria	50	50	50	1	1	0	0	16.9	11.0
Marche	150	150	148	1	0.999821	0	0.00016970	17.3	11.3
Latium	80	79	79	0.999684	0.999684	0.00030111	0.00030111	17.6	11.8
Abruzzo	24	24	24	1	1	0	0	17.9	11.5
Apulia	83	83	83	1	1	0	0	17.3	11.4
Calabria	50	50	49	1	0.999184	0	0.00075415	15.7	10.2
Sicily	304	301	297	0.999935	0.999848	0.00006262	0.00013863	17.6	11.5
Rural areas	1046	1043	1032	0.999995	0.999974	0.00000543	0.00002428	17.4	11.3
Urban areas	463	462	458	0.999991	0.999953	0.00000935	0.00004477	17.4	11.4
Total	1509	1505	1490	0.999996	0.999983	0.00000348	0.00001588	17.4	11.3

Table 1 - Standard diversity indexes for 13 and 11 RM Y-STRs in the total set of 1509 Italian individuals, summarized according to regional populations and rural/urban origins.

		Haplotypes				*h*				θ			
	n	13 RM Y-STR	11 RM Y-STR	Yfiler	PPY23	13 RM Y-STR	11 RM Y-STR	Yfiler	PPY23	13 RM Y-STR	11 RM Y-STR	Yfiler	PPY23
Total	1133	1129	1117	1087	-	0.999994	0.999975	0.999914	-	0.00000615	0.00002358	0.000113238	-
Rural	846	843	833	812	-	0.999992	0.999964	0.999888	-	0.00000827	0.00003420	0.000146265	-
Urban	287	286	284	279	-	0.999976	0.999927	0.999781	-	0.00002403	0.00007011	0.000262250	-
Total	938	935	923	-	918	0.999991	0.999966	-	0.999950	0.00000844	0.00003202	-	0.00005862
Rural	685	683	673	-	670	0.999991	0.999949	-	0.999932	0.00000844	0.00004776	-	0.00007444
Urban	253	252	250	-	250	0.999969	0.999906	-	0.999906	0.00003088	0.00008976	-	0.00008976

Table 2 - Comparison of standard diversity indexes in a subset of Italian individuals included in the RM Y-STR study who were also genotyped with the Yfiler (n=1133) and PPY23 (n=938) systems.

		AMOVA		
		Within populations	Among populations within groups	Among groups
13 RM Y-STR	Total (n=1509)	99.996%	0.003%	0.001%
	Rural (n=1046)	99.993%	0.008%	-0.001%
	Urban (n=463)	99.996%	0.003%	0.001%
11 RM Y-STR	Total (n=1509)	99.985%	0.015%	0.000%
	Rural (n=1046)	99.978%	0.022%	0.000%
	Urban (n=463)	99.986%	0.011%	0.003%
YFiler	Total (n=1133)	99.947%	0.048%	0.005%
	Rural (n=846)	99.904%	0.102%	-0.006%
	Urban (n=287)	99.954%	0.056%	-0.010%
PPY23	Total (n=938)	99.976%	0.019%	0.005%
	Rural (n=685)	99.965%	0.031%	0.004%
	Urban (n=253)	99.981%	0.028%	-0.009%

Table 3 - AMOVA results for different Y-STR panels in the total set of Italian individuals and rural/urban subsamples.